

## INTERNET CRAWL SEEDING

### CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority from U.S. provisional application number 60/252,273, filed on November 21, 2000, which is herein incorporated  
5 by reference in its entirety. This application is related to the following applications, each being filed on the same date as the instant application: application number \_\_\_\_\_, entitled "Internet Streaming Media Workflow Architecture," attorney docket number D4961-00014; application number \_\_\_\_\_, entitled "Interpretive Stream Metadata Extraction," attorney docket  
10 number D4961-00015; application number \_\_\_\_\_, entitled "Metadata Quality Improvement," attorney docket number D4961-00016; application number \_\_\_\_\_, entitled "Full Text Relevancy Ranking," attorney docket number D4961-00017; application number \_\_\_\_\_, entitled "Grouping Multimedia And Streaming Media Search Results," attorney docket number D4961-00018;  
15 and application number \_\_\_\_\_, entitled "Fuzzy Database Retrieval," attorney docket number D4961-00019.

### FIELD OF THE INVENTION

[0002] The present invention relates to computer related information search and retrieval, and specifically to providing search terms to multimedia  
20 and streaming media search tools.

### BACKGROUND

[0003] An aspect of the Internet (also referred to as the World Wide Web, or Web) that has contributed to its popularity is the plethora of multimedia and streaming media files available to users. However, finding a  
25 specific multimedia or streaming media file buried among the millions of files

on the Web is often an extremely difficult task. The volume and variety of informational content available on the web is likely continue to increase at a rather substantial pace. This growth, combined with the highly decentralized nature of the web, creates substantial difficulty in locating particular informational content.

5

**[0004]** Streaming media refers to audio, video, multimedia, textual, and interactive data files that are delivered to a user's computer via the Internet or other network environment and begin to play on the user's computer before delivery of the entire file is completed. One advantage of streaming media is that streaming media files begin to play before the entire file is downloaded, saving users the long wait typically associated with downloading the entire file. Digitally recorded music, movies, trailers, news reports, radio broadcasts and live events have all contributed to an increase in streaming content on the Web. In addition, less expensive high-bandwidth connections such as cable, DSL and T1 are providing Internet users with speedier, more reliable access to streaming media content from news organizations, Hollywood studios, independent producers, record labels and even home users themselves.

10

15

**[0005]** A user typically uses a search engine to find specific information on the Internet. A search engine is a set of programs accessible at a network site within a network, for example a local area network (LAN) or the Internet and World Wide Web. One program, called a "robot" or "spider", pre-traverses a network in search of documents (e.g., web pages) and builds large index files of keywords found in the documents. Typically, a user formulates a query comprising one or more search terms and submits the query to another program of the search engine. In response, the search engine inspects its own index files and displays a list of documents that match the search query, typically as

20

25

hyperlinks. The user then typically activates one of the hyperlinks to see the information contained in the document.

[0006] Search engines, however, have drawbacks. For example, many typical search engines are oriented to discover textual information only. In particular, they are not well suited for indexing information contained in structured databases (e.g. relational databases), voice related information, audio related information, multimedia, and streaming media, etc. Also, mixing data from incompatible data sources is difficult for conventional search engines.

[0007] Another disadvantage of conventional search engines is that irrelevant information is aggregated with relevant information. For example, it is not uncommon for a search engine on the web to locate hundreds of thousands of documents in response to a single query. Many of those documents are found because they coincidentally include the same keyword in the search query. Sifting through search results in the thousands, however, is a daunting task. For example, if a user were looking for a song having the title "I Am The Walrus," the search query would typically contain the word "walrus." The list of hits would include documents providing biological information on walruses, etc. Thus, the user would have to review an enormous number of these hits before finally (if ever) reaching a hit related to the desired song title. Adding to a user's frustration is the possibility that many of the search results are duplicates and/or variants of each other, leading to the same document (e.g. uniform resource locator, URL). Further difficulty occurs in trying to evaluate the relative merit or relevance of concurrently found documents. The search for specific content based on a few key words will almost always identify documents whose individual relevance is highly variable.

[0008] Thus, there is a need for an automated media search tool that provides information to a user that overcomes the previously described drawbacks and disadvantages.

### SUMMARY OF THE INVENTION

5 [0009] A method and system for seeding a search system for searching for media on a computer network includes providing at least one search term to the search system. At least one search result is received. The search result(s) includes metadata associated with the media. The search result(s) are parsed. The parsed metadata is provided as a new search term to the search system.

10

### BRIEF DESCRIPTION OF THE DRAWINGS

[0010] The above and other advantages and features of the present invention will be better understood from the following detailed description of the preferred embodiments of the invention, which is provided in connection with the accompanying drawings. The various features of the drawings may not be to scale. Included in the drawing are the following figures:

15

[0011] Figure 1 is a block diagram of a computer system in accordance with an exemplary embodiment of the present invention;

[0012] Figure 2 is a flow diagram of an exemplary search and retrieval process in accordance with the present invention;

20

[0013] Figure 3 is a functional block diagram of an exemplary multimedia and/or streaming media metadata search, retrieval, and enhancement system in accordance with the present invention;

[0014] Figure 4 is a flow diagram of an exemplary spider seeding process in accordance with the present invention;

[0015] Figure 5 is a flow diagram of an exemplary distribution and extraction process in accordance with the present invention;

5 [0016] Figure 6 is a flow diagram of an exemplary interpretive metadata extraction and database retrieval process in accordance with the present invention;

[0017] Figure 7 is a flow diagram of an exemplary process for querying databases in accordance with the present invention;

10 [0018] Figure 8 is a flow diagram of an exemplary grouping process in accordance with the present invention;

[0019] Figure 9 is a flow diagram of an exemplary iterative masking process in accordance with the present invention;

15 [0020] Figure 10 is a flow diagram of an exemplary metadata quality improvement process in accordance with the present invention; and

[0021] Figure 11 is a flow diagram of an exemplary full-text relevancy ranking process in accordance with the present invention.

### DETAILED DESCRIPTION

20 [0022] Although the invention is described in terms of exemplary embodiments, it is not limited thereto. Rather, the appended claims should be construed broadly, to include other variants and embodiments of the invention, which may be made by those skilled in the art without departing from the scope and range of equivalents of the invention.

[0023] The present invention is a system and method for retrieving media files and data related to media files on a computer network via a search system utilizing metadata. As used herein, the term “media file” includes audio, video, textual, multimedia data files, and streaming media files.

5 Multimedia files comprise any combination of text, image, video, and audio data. Streaming media comprises audio, video, multimedia, textual, and interactive data files that are delivered to a user’s computer via the Internet or other communications network environment and begin to play on the user’s computer/ device before delivery of the entire file is completed. One  
10 advantage of streaming media is that streaming media files begin to play before the entire file is downloaded, saving users the long wait typically associated with downloading the entire file. Digitally recorded music, movies, trailers, news reports, radio broadcasts and live events have all contributed to an increase in streaming content on the Web. In addition, the reduction in cost of  
15 communications networks through the use of high-bandwidth connections such as cable, DSL, T1 lines and wireless networks (e.g., 2.5G or 3G based cellular networks) are providing Internet users with speedier, more reliable access to streaming media content from news organizations, Hollywood studios, independent producers, record labels and even home users themselves.

20 [0024] Examples of streaming media include songs, political speeches, news broadcasts, movie trailers, live broadcasts, radio broadcasts, financial conference calls, live concerts, web-cam footage, and other special events. Streaming media is encoded in various formats including REALAUDIO<sup>®</sup>, REALVIDEO<sup>®</sup>, REALMEDIA<sup>®</sup>, APPLE QUICKTIME<sup>®</sup>, MICROSOFT  
25 WINDOWS<sup>®</sup> MEDIA FORMAT, QUICKTIME<sup>®</sup>, MPEG-2 LAYER III AUDIO, and MP3<sup>®</sup>. Typically, media files are designated with extensions (suffixes) indicating compatibility with specific formats. For example, media files (e.g., audio and video files) ending in one of the extensions, .ram, .rm,

.rpm, are compatible with the REALMEDIA<sup>®</sup> format. Some examples of file extensions and their compatible formats are listed in the following table. A more exhaustive list of media types, extensions and compatible formats may be found at <http://www.bowers.cc/extensions2.htm>.

Format	Extension
REALMEDIA <sup>®</sup>	.ram, .rm, .rpm
APPLE QUICKTIME <sup>®</sup>	.mov, .qif
MICROSOFT WINDOWS <sup>®</sup> MEDIA PLAYER	.wma, .cmr, .avi
MACROMEDIA FLASH	.swf, .swl
MPEG	.mpg, .mpa, .mpl, .mp2
MPEG-2 LAYER III Audio	.mp3, .m3a, .m3u

5

[0025] Metadata, literally means “data about data.” Metadata is data that comprises information that describes the contents or attributes of other data (e.g., media file). For example, a document entitled, “Dublin Core Metadata for Resource Discovery,” (<http://www.ietf.org/rfc/rfc2413.txt>) separates metadata into three groups, which roughly indicate the class or scope of information contained therein. These three groups are: (1) elements related primarily to the content of the resource, (2) elements related primarily to the resource when viewed as intellectual property, and (3) elements related primarily to the instantiation of the resource. Examples of metadata falling into these groups are shown in the following table.

10

15

Content	Intellectual Property	Instantiation
Title	Creator	Date
Subject	Publisher	Format
Description	Contributor	Identifier
Type	Rights	Language
Source		
Relation		
Coverage		

[0026] Sources of metadata include web page content, uniform resource locators (URLs), media files, and transport streams used to transmit media files. Web page content includes HTML, XML, metatags, and any other text on the web page. As explained in more detail, herein, metadata may also be  
5 obtained from the URLs the web page, media files, and other metadata.

Metadata within the media file may include information contained in the media file, such as in a header or trailer, of a multimedia or streaming file, for example. Metadata may also be obtained from the media/ metadata transport stream, such as TCP/IP (e.g., packets), ATM, frame relay, cellular based  
10 transport schemes (e.g., cellular based telephone schemes), MPEG transport, HDTV broadcast, and wireless based transport, for example. Metadata may also be transmitted in a stream in parallel or as part of the stream used to transmit a media file (a High Definition television broadcast is transmitted on one stream and metadata, in the form of an electronic programming guide, is  
15 transmitted on a second stream).

[0027] Figure 1 is a block diagram illustrating a system, generally designated 100, in accordance with an exemplary embodiment of the present invention. The system 100 includes a plurality of server computers 18, 20, a plurality of user computers 12, 14, and a plurality of databases 21, 22. The  
20 server computers 18, 20 and the user computers 12, 14 may be connected by a network 16, such as for example, an Intranet or the Internet. The user computers 12, 14 may be connected to the Intranet or Internet by a modem connection, a Local Area Network (LAN), cable modem, digital subscriber line (DSL), or other equivalent coupling means. Alternatively, the computers  
25 communicate through a communications network by a cable, twisted pair, wireless based interface (cellular, infrared, radio waves) or equivalent connection utilizing data signals. Databases 21, 22 may be connected to the user computers and the server computers by any means known in the art.



Databases may take the form of any appropriate type of memory (e.g., magnetic, optical, etc.). Databases 21, 22 may be external memory or located within the server computer or the user computer. Each user computer 12, 14 preferably includes a video display device for displaying information and a browser program (e.g. MICROSOFT INTERNET EXPLORER<sup>®</sup>, NETSCAPE NAVIGATOR<sup>®</sup>, etc.), as is well known in the art.

**[0028]** Computers may also encompass computers embedded within consumer products and other computers. For example, an embodiment of the present invention may comprise computers (as a processor) embedded within a television, a set top box, an audio/video receiver, a CD player, a VCR, a DVD player, a multimedia enable device (e.g., telephone), and an Internet enabled device.

**[0029]** In an exemplary embodiment of the invention, the server computers 18, 20 include one or more program modules and one or more databases which allow the user computers 12, 14 to communicate with the server computer, and each other, over the network 16. The program module(s) of the server computers 18, 20 include program code, written in PERL, Extensible Markup Language (XML), Java, Hypertext Mark-up Language (HTML), or any other equivalent language which allows the user computers 12, 14 to access the program module(s) of the server computer through the browser programs stored on the user computers. Although only two user computers 12, 14, two server computers 18, 20, and two databases 21, 22 are labeled in Figure 1, those of ordinary skill in the art will realize that the system 100 may include any number of user computers, server computers, and databases.

**[0030]** In an exemplary embodiment of the present invention, media files and related metadata are searched for and retrieved by reading, extracting,

enhancing, and grouping metadata describing the contents of files. Figure 2 is a flow diagram of an exemplary search and retrieval process in accordance with the present invention. Discovery (step 24) comprises an automated process referred to as a spider or web crawler, for searching web sites or data available through a communications network. Each web site may comprise any number of web pages and/or data on storage devices (hard drives, flash cards, disc drives, optical disc storage). The spider utilizes predetermined algorithms to continuously search for media files on web pages and file directories at each searched web site. The spider also searches each web site for links to other web sites, unique streams, and downloadable files.

**[0031]** Upon finding a media file, metadata associated with that file is extracted (step 26). Metadata is extracted from sources such as the name of the media file, the MIME responses, links to the media file, text surrounding the media file on the website, metatags (descriptive information embedded in sources as program code or HTML) in or surround the media file, content partners supplying metadata about their files, and the results of reading the metadata of the media file with an interpretive extraction process.

**[0032]** Extracted metadata is enhanced in step 28. The extracted metadata associated the media files are stored in memory (e.g., transferred to a database). The metadata is assessed, analyzed, and organized in accordance with attributes associated with the media file. If metadata information is missing from the extracted metadata, it is added (step 28). If metadata information is incorrect, it is corrected (step 28). For example, if metadata associated with a song comprises the fields of Composer, Title, Musician, Album Name, and Music Genre, but is missing the date the song was copyrighted, the copyright date is added to the extracted metadata. The metadata (e.g., copyright date) used to enhance the extracted metadata is

obtained from at least one of several sources. These sources include a baseline database of metadata associated with the search target (e.g., the particular song of interest) and the semantic and technical relationships between the fields in the extracted metadata.

- 5 [0033] The extracted metadata, which may be enhanced, is categorized in accordance with specific metadata attributes in step 30. At this point the links, e.g., uniform resource indicators (URIs) in the form of uniform resource locators (URLs) for web pages and data files, may be transferred to the user, the URL of the media file may be transferred to the user, or the categorized
- 10 metadata may be used (e.g., transferred to a search engine) to search and retrieve the target media file. In an exemplary embodiment of the invention, the target streaming media stream automatically starts playing. For example, a specific song is searched for, and the ultimate result is the playing of the song on the user's computer system.
- 15 [0034] Uniform resource indicators (URIs) are a universal set of names that refer to existing protocols or name spaces that identify resources (e.g., website, streaming media server,), services (e.g., videos on demand, internet radio), devices (e.g., mobile phone, internet enable appliance), and data files (e.g., media files and text documents). A URL is a form of a URI that expresses
- 20 an address that maps to an access algorithm using network protocols (e.g., TCP/IP or a MPEG transport scheme). When a URL is used, a specific resource, service, device, or data file may be accessed and/or manipulated. An alternative form of a URI known as an Internet protocol number or address (IP) is a series of numbers that refers to a specific resource, service, or data file.
- 25 Optionally, a URL is mapped to an IP number, which provides two ways to access a desired resource (e.g., a resource is accessed either by using [www.whitehouse.gov](http://www.whitehouse.gov) or the IP address 198.137.240.91).

**[0035]** Figure 3 is a functional block diagram of an exemplary search and retrieval system, designated 300, in accordance with the present invention. System 300 comprises a plurality of autonomous, interacting agents for  
5 collecting, extracting, enhancing, and grouping media metadata. Although system 300 depicts the agents performing in an exemplary order, agents may perform respective functions in any order. Each agent receives and provides data from and to data queues. Data residing on a data queue is available to all agents. In an exemplary embodiment of the invention, media files and  
10 associated metadata are stored in memory (e.g., a database) and assigned an identifier (id). The ids are enqueued, and the agents receive and provide the ids from and to the queues. Agents retrieve associated data (e.g., metadata) from memory to perform various functions, and store the processed data in memory (e.g., update the database).

**[0036]** Spider 66 incorporates a process of seeding to search for media and related files. Figure 4 is a flow diagram of an exemplary spider seeding process in accordance with the present invention. The spider is seeded in step 36. The spider seeds its search by adding terms that are related to the query being used to index media. Additionally, the spider adds media related terms  
20 to the search, such as "MP3" and "Real Audio". Adding media related terms to the search tend to limit the search to media related files and URIs (in the form of links). For example, adding streaming media related terms to the search tends to limit the search to streaming media related files and links. The spider receives the search results and uses the links to perform more searches. The  
25 input queue of the spider may be seeded with several types of information, such as the results of querying other search engines, manually generated sets of web page URLs, and processing proxy cache logs (i.e., web sites that other users have accessed).

**[0037]** The spider uses seed URLs to search (step 38) and retrieve (step 40) the HTML text from located web sites. The file name and MIME type of the web site are stored in memory. The text is parsed to look for links to other web resources associated with media in step 42. The HTML code of each web page is examined for anchor tags, form elements, known JavaScript functions, etc., to find other web resources associated with media. These newly found web resources are used as seeds for the spider for additional searches (added to the spider input queue) by repeating steps 36 through 42 using the new seeds.

**[0038]** Referring again to Figure 3, the parsed results (from step 42 in Figure 4) relating to the media are passed to extraction agent 68 via an extraction queue 67. Results not associated with the media are not pursued. The extraction queue 67 comprises URLs to be analyzed with respect to associated media metadata. The extraction queue 67 may comprise metadata queue entries such as media URLs, Web page URLs, Web page titles, Web page keywords, Web page descriptions, media title, media author, and media genre. Each queue entry added to the extraction queue is assigned a processing time and a priority. In an exemplary embodiment of the invention, each queue entry is given a processing time of “now” and the same default priority. The iterative seeding process increases the number of queue entries added to the extraction queue 67.

**[0039]** The extraction agent 68 comprises an interpretive metadata extractor and a database retriever. The extraction agent 68 distributes and performs enhanced metadata extraction of queue entries on the extraction queue 67. Figure 5 is a flow diagram of an exemplary distribution and extraction process in accordance with the present invention. Queue entries contained in the extraction queue 67 are dequeued and distributed to media specific extractors in step 46. The extraction queue entries are dequeued and

distributed in priority and time order. Preferably, the file extension, MIME type, and/or file identification for each queue entry is examined to determine the type of media format. The queue entry is then sent to the appropriate media specific extractor. Optionally, other types of data are used to determine the media format of a file (for example, the extraction queue 67 reads the metadata embedded in a media file to determine that it is a Real Media video file).

**[0040]** In step 48, queue entries sent to a specific media extractor are extracted by that specific extractor. Metadata extraction comprises the process of extracting metadata from the media file or from related media content (e.g., from the referring web page). Types of media specific extractors include multimedia and streaming extractors that can extract metadata from formats such as REALAUDIO<sup>®</sup>, REALVIDEO<sup>®</sup>, REALMEDIA<sup>®</sup>, APPLE QUICKTIME<sup>®</sup>, MICROSOFT WINDOWS<sup>®</sup> MEDIA FORMAT, QUICKTIME<sup>®</sup>, MPEG-2 LAYER III AUDIO, and MP3<sup>®</sup>, for example. As discussed in more detail herein, interpretive metadata extraction captures and aggregates metadata pertaining to a media file using metadata from the media stream, third party databases, the referring web page, and the URL, and replaces inaccurate metadata with known good metadata. An Internet stream object is created comprising the media file from the URL, metadata extracted from the media file and an identifier (id). The Internet stream object is automatically stored in memory (step 50). In an exemplary embodiment of the invention, memory storage comprises providing the object to a relational database management system (DBMS) for storage and database management.

**[0041]** In step 52, it is determined if the accessible media file and the associated metadata links are valid. Validation comprises determining if the Web page comprises a link to a desired media file, and also determining if the desired media file works. In an exemplary embodiment of the invention, a

streaming media file is retrieved and played to determine it is valid. If determined to be invalid (not successful in step 52), the Internet stream object is assigned a later time and priority. The Internet stream object is queued to the extractor, and steps 48 through 50 are repeated with at the later time and in accordance with the newly assigned priority. If extraction is valid (successful in step 52), the Internet streaming object is queued and available to all agents.

**[0042]** Extraction agent 68 captures and aggregates media specific metadata pertaining to the media (including multimedia and streaming media) from sources such as the media URL, the referring Web page URL, title, key words, description, and third party databases. Figure 6 is a flow diagram of an exemplary interpretive metadata extraction and database retrieval process in accordance with the present invention. Metadata, which may be inaccurate or “noisy,” is extracted (step 60), parsed and indexed (step 62), compared with fields in known databases (step 64), and replaced (step 65) with accurate metadata obtained from a valid (ground truth) database. Metadata is indexed and parsed into metadata fields (step 62) and compared to other databases (step 64), such as a music or movie database, whose accuracy is known (ground truth databases). Ground truth databases may be obtained from sources such as third party databases, previously created databases, and user entered databases, for example. Noisy fields are corrected and/or replaced with accurate data (step 65). New fields are added if appropriate (step 65).

**[0043]** For example, assume the spider 66 finds a media file containing a music song. The metadata is extracted by extracting agent 68, and parsed and indexed into the following metadata fields: the referring URL, the media URL, the title, and the performer of the song. The information contained in these fields is as follows.

FIELD	CONTENTS
The referring URL	www.singingfish.com/index.html
Media URL	www.singingfish.com/foobar.RAW
Title	"I am the Fishman"
Performer	Paul McCarpney

[0044] The metadata fields are compared to a known database, such as a third party database, to compare contents of the metadata fields with the contents of the fields in the known database. In this example, assume a known database is located and contains the following indexed metadata.

FIELD	CONTENTS
Copyright	1984
Title	"We are the Fishmen"
Album	Rubber Suit
Music Genre	Light Rock
Performer	John Lennon
Performer	Paul McCarpney

[0045] The interpretive extraction agent 68, adds the missing fields corresponding to the copyright, album, music genre, and composer, and adds the additional performer (i.e., John Lennon). The interpretive extraction 68 also corrects the title of the song from "I am the Fishman" to "We are the



Fishmen” because the database comprises valid or authoritative metadata. Thus, prior to these enhancements, this media file could only be located if a user enter “Paul McCarpney” as the performer and/or “I am the Fishman” as the title. As a result of the enhancements provided by the interpretive metadata extraction agent 68, a user may locate this media file also by searching for any of the resultant fields (e.g., the album name or the composer).

**[0046]** Not all databases queried are determined to be ground truth databases. Figure 7 is a flow diagram of an exemplary process for querying databases in accordance with the present invention. Noisy metadata (metadata that needs to be cleaned up before being processed) are compared to potential ground truth databases to determine if a potential ground truth database qualifies as a ground truth database. In step 84, noisy metadata in each field (e.g., artist, title, album) is separated into keywords by transforming any connecting characters into white space. For example, “oops\_i\_did\_it\_again” is transformed to the cleaned up “oops i did it again”. Connecting characters may include, for example, period (“.”), underscore (“\_”), backslash (“\”), forward slash (“/”), comma (“,”), asterisk (“\*”), hyphen (“-“), and/or any other appropriate connecting character. The fields in the noisy metadata are used to perform a full-text query against one or more fields in the potential ground truth databases (step 86).

**[0047]** A score is calculated, in step 88, to quantify the degree of similarity between the noisy data (candidate metadata) and potential ground truth data (valid metadata). In an exemplary embodiment of the invention, the number of matching keywords in the fields being compared determines a score. For example, if the input query is “oops i did it” and two potential ground truth data records are “oops i did it again” and “did it again for you”, the first score is 4 and the second score is 2. In an alternate embodiment of the invention,

fields are also assigned weights, which are multiplied by the number of matching keywords. For example, the artist field may be assigned a weight of 3, and the copyright date field may be assigned a weight of 1. Thus, if two keywords match in each of the artist and copyright fields, the score for the artist field is 6, and the score for the copyright field is 2. Further, individual field scores may be added, averaged, or combined by any appropriate means to derive a cumulative database score. The scores are compared to a threshold value (step 90) to determine if the potential ground truth database qualifies (step 92), or does not qualify (step 94) as a ground truth database. If a database qualifies as a ground truth database, it is used by the interpretive extraction process as described herein. The threshold value may be predetermined and constant, or may be adaptively determined in accordance with the range of calculated scores.

**[0048]** Referring again to Figure 3, the validator 72 dequeues entries from the queue in time and priority order. The validator 72 validates the media data by determining if the Web page comprises a link to a desired media file and also determining if the desired media file works. Validation is performed at a future point in time (e.g., check if the URL is still alive in 3 days), or alternatively, at periodic future points in time. If validity changes from valid to invalid, a notification is sent to promoter 82, as indicated by arrow 70. Validity may change from valid to invalid, for example, if the media file was removed from the linking URL.

**[0049]** The virtual domain detector 74 dequeues data from the queue in time and priority order. The virtual domain detector 74 looks for duplicate domains (field of the URL). If duplicates are found, they are identified as such and queued. The queued ids are available to all agents.

5 [0050] It is not uncommon for Web pages and multiple servers with  
different portions of a URL, e.g., different domains, to host the same media  
content. Further, the same media content may be available in different formats  
and bit rates. The grouper 76 analyzes and compares URLs in the database.  
10 The grouper 76 combines variants of the same media URL and creates a group  
in which all metadata for similar URLs are added to the group. URLs are  
analyzed to determine if they are variations of related files. For example, if  
two URLs share a very complex path that differs only in the file extension, the  
two URLs are considered to be related. Differences are eliminated by masking  
out tokens at the same relative location with respect to the original string.

15 [0051] Figure 8 is a flow diagram of an exemplary grouping process in  
accordance with the present invention. Grouping comprises the steps of  
binning 102 and iterative masking 104. Binning 102 comprises the steps of  
selecting and sorting URLs (step 106) and combining URLs having common  
specified attributes into bins (step 108). In step 106, each URL in the database  
is analyzed to determine the contents of specific fields. URLs having similar  
contents in the specified fields are placed (binned) into common sets or bins of  
URLs (step 108). All URLs in the database are placed into bins. Each bin has  
a smaller number of URLs than the number of URLs in the database.  
20 Although, it is possible that all URLs in the database are placed into the same  
bin, it is highly unlikely. As a result of the binning process 102, each bin  
comprises at least one URL, and the URLs contained in bins comprising a  
plurality of URLs have at least one common attribute (i.e., same content in  
specified field(s)). Examples of specified fields include fields indicating artist,  
25 linking URL, title, copyright, host URL, duration, bit rate, sampling rate, etc.  
In an exemplary embodiment of the invention, URLs are binned if they have  
the same content for the fields indicating host URL and duration. One  
advantage of binning is that the number of URLs contained in a bin is relatively

small compare to the number of URLs contained in the database, thus partitioning the URLs in the database into more manageable sets of URLs.

5 [0052] Selected bins are iteratively masked in step 104. The masking process 104 is performed on URLs on a bin by bin basis. Each field of each URL is compared to a mask. Not all bins require processing by the iterative masking process 104. In an exemplary embodiment of the invention, bins containing only a single URL are not iteratively masked 104, and bins containing a plurality of URLs are processed in accordance with the iterative masking process 104.

10 [0053] Figure 9 is a flow diagram of an exemplary iterative masking process in accordance with the present invention. Iterative masking (step 104) comprises creating a “mask” (step 110) and comparing the mask with each URL in a bin (step 112). A mask comprises at least one character to be removed from the contents of a field within a URL. In an exemplary  
15 embodiment of the invention, a mask is a string of characters. For example, a mask may comprise a string of characters pertaining to bit rate of the streaming media content, formatting of the streaming media, or any related characteristic. The mask is compared to each field in a URL in a bin, in step 112. It is determined if any of the characters in the mask match characters in the URL  
20 (step 114). If a match exists, the matching character, or characters, is removed from the URL (step 116), otherwise the URL is unchanged. This process is repeated until all URLs in the bin have been compared with the mask (step 118).

25 [0054] Resultant URLs (i.e., URLs that have been compared to the mask) in the bin are compared and collapsed into a single URL if they are the same (step 120). For example, if four URLs differ only by bit rate, and the bit rate of each URL is masked out, the resulting four URLs are collapsed into a

single URL. If more bins have been selected, the iterative masking process is repeated for the next bin (step 122) starting at step 112. Grouped URLs are queued and available for all agents.

[0055] For example, assume all URLs in the database have been binned such  
5 that all URLs comprising the same referring URL are binned together. Thus,  
assume the following URLs are in the same bin.

http://foo.bar.com/video/someArtist/myVideo\_28.ram

http://foo.bar.com/video/someArtist/myVideo\_56.ram

http://blatz.com/56/someArtist/yourVideo.ram

10 http://blatz.com/28/someArtist/yourVideo.ram

Further assume that the mask is a string of characters related to bit rate including 28, 56, and 100. Applying this mask to the above URLs and removing the matched characters results in the following URLs.

http://foo.bar.com/video/someArtist/myVideo\_.ram

15 http://foo.bar.com/video/someArtist/myVideo\_.ram

http://blatz.com//someArtist/yourVideo.ram

http://blatz.com//someArtist/yourVideo.ram

20 Instead of the bin containing four unique URLs, the bin now contains two copies each of two unique URLs. Each of the two copies is collapsed into a single URL, resulting in the following URLs.

http://foo.bar.com/video/someArtist/myVideo\_.ram

http://blatz.com//someArtist/yourVideo.ram

[0056] Referring again to Figure 3, metadata quality improver 78 dequeues entries in time and priority order. Metadata quality improver 78

enhances metadata by adding fields of metadata based upon the contents of the fields in the URL of the media file and the contents of the fields in the URL of the referring Web page. The media file is then searchable under the subject heading of the added metadata. For example, a streaming media file may have a referring Web page at www.cnn.com. The metadata quality improver 78 adds the term “news” to the metadata associated with the streaming media file, because cnn is related to news. As a result, the streaming media file is now searchable under the subject heading of “news”. Expert based rules are used to associate field contents with metadata. Metadata quality improver 78 applies rules to eliminate duplicate URLs that point to the same data, rules to collect variants of media files with the same content but different encodings or formats (e.g., for multimedia and streaming media), and rules to update metadata fields using prefix URL associations. The metadata quality improvement process comprises prefix rule evaluation, genre annotation, and MUZE<sup>®</sup> (a commercial database containing metadata about music including song title, music author, and album information) annotation.

**[0057]** Figure 10 is a flow diagram of an exemplary metadata quality improvement process comprising prefix rule evaluation, genre annotation, and MUZE<sup>®</sup> annotation in accordance with the present invention. Prefix rule evaluation comprises reorganizing the fields in the media URL and determining if an association exists between known sets of metadata and the first field content. Genre annotation comprises updating the genre metadata to ensure proper formatting. MUZE<sup>®</sup> annotation comprises editing fields of the metadata to improve the quality of other fields of the metadata.

**[0058]** The fields of the URL are reorganized in step 138. In an exemplary embodiment of the invention, the URL is reorganized in reverse order. Thus the first field of the URL becomes the last field and the last field

becomes the first. In many instances, this results in a reorganized URL having its most specific field first and its least specific field last. In many instances, this also results in the first set of contiguous fields (i.e., prefix) of the reorganized URL having associated metadata. The first field of the reorganized

5 URL is analyzed to determine if an association exists between the first field and predetermined sets of metadata (step 140). Predetermined sets of metadata may comprise metadata obtained from other fields in the metadata and/or terms (metadata) contained in a database of terms. If it is determined that an association exists (step 142), the associated metadata are added to the original

10 metadata in step 148. After metadata are added, it is determined if the reorganized URL contains more fields (which have not been analyzed for associated metadata) in step 150. If no associated metadata are identified (step 140 and 142), it is also determined if more fields exist (step 150). If more fields exist, the next field is analyzed to determine if an association exists

15 between the next field and the predetermined sets of metadata (step 146). In an exemplary embodiment of the invention, the next field is the next contiguous field. If no associated metadata are identified (step 142), no new metadata are added to the metadata associated with the media file. If associated metadata are identified, the associated metadata are added to the original metadata in

20 step 148. This process is continued until all the fields in the reorganized URL have been analyzed. At this point, metadata associated with the longest match (i.e., the greatest number of fields having associated metadata) have been added to the original metadata. Databases are updated with the newly added metadata, and the associated ids are queued and available to all agents.

25 **[0059]** In an exemplary embodiment of the invention, the genre metadata is updated to ensure proper formatting and correctness. The updated metadata is analyzed to determine if the genre field(s) are correct. If it is determined that the genre field(s) are not correct, they are updated. The genre fields are

updated in accordance with predetermined association rules. For example, assume the contents of a field pertaining to category is “music” (i.e., “category=music”). The metadata is analyzed, and the metadata terms “artist=Freddy Roulette”, “genre\_MP3=punk rock” are found. The field associated with category will be changed from music to punk rock, resulting in “category=punk rock”. In this example, the category field is changed because a predetermined association rule is encoded to change the “category” field to the same as the “genre\_MP3” field.

**[0060]** In another exemplary embodiment of the invention, the iterative process is halted after metadata associated with the longest prefix of contiguous fields of the reorganized URL are identified, and metadata associated with the prefix, and not the individual fields is added to the original metadata. For example, assuming a URL has ten fields, if the first four fields of the reorganized URL have associated metadata, and the fifth field does not have associated metadata, the sixth through the tenth fields are not analyzed for associated metadata. In this example, the metadata associated with the first four fields, i.e., the prefix, and not the individual fields, (and, as will be explained herein with reference to muze annotation, possibly the metadata associated with the fifth field) are added to the original metadata.

**[0061]** Metadata is added to the metadata associated with the media file. Added metadata may comprise metadata corresponding to category, title, delivery mode, publisher, performer, program title, creation country, and language, for example. The added metadata may be in the form of textual data (e.g., new terms) and/or URLs (e.g., new links). Also, in accordance with the MUZE<sup>®</sup> annotation rule, added metadata may comprise the content of the field in the reorganized URL that is next to the matching prefix (e.g., first non-matching field). The content of the field is edited to replace connecting



characters with spaces, and then added as new metadata. Connecting characters may include, for example, period (“.”), underscore (“\_”), backslash (“\”), forward slash (“/”), comma (“,”), asterisk (“\*”), hyphen (“-“), and/or any other appropriate connecting character. This muze annotation rule is

5 advantageous for URLs comprising field content of “MP3”. In an exemplary embodiment of the invention, all reorganized URLs beginning with the prefix “com.MP3.downloads” are categorized as music and the recommended title (i.e., song title) is based on the filename as given in the field next to the matched prefix. For example, assume the following reorganized URL,

10 “com.MP3.downloads/Freddy Roulette/Laundry\_Mat\_Blues”. In this example, Freddy Roulette is the content of the artist field and Laundry\_Mat\_Blues is the content of the title field. The metadata quality improvement process finds a match for the prefix of the reorganized URL ending with Freddy Roulette. Because the reorganized URL begins with com.MP3, the metadata quality

15 improver 78 edits the next field (i.e., Laundry\_Mat\_Blues) after the matched prefix and adds the edited data as the recommended title of the song. The edited field content has no underscores. Thus the resultant added metadata in this example is “Laundry Mat Blues”. Databases are updated with the newly added metadata, and the associated ids are queued and available to all agents.

20 Examples of the types of metadata that are added to matched fields are shown in the following table.

Field Prefix	Added Metadata
org.npr.www/ramfiles/atc	Category: Radio Delivery Mode: Broadcast Publisher: NPR Performer: Noah Adams Program Title: All Things Considered Language: English
com.sportsline.www/u/audio/basketball/nba	Category: Sports Genre: Basketball Creation Country: US Language: English
com.msnbc.www	Category: News

	Recommended Title: Referring Page Title
com.mp3	Category: Music
com.mp3.downloads	Category: Music Recommended Title: Filename in the next field of the URL (i.e., text after the matched prefix)

[0062] The full-text relevancy ranker 80 comprises ranking and sorting data (e.g., media metadata) based on a variety of semantic and technical data fields (e.g., title, author, date, duration, bit rate, etc.). Full-text relevancy ranker 80 is depicted as part of the work flow architecture of system 300. This depiction is exemplary. In another embodiment of the invention, full-text relevancy ranker 80 is not part of the workflow architecture. The option to include full-text relevancy ranker 80 as part of the workflow architecture (or not) is depicted by the dashed arrows in Figure 3 (from metadata quality improver 78 to full-text relevancy ranker 80, from full-text relevancy ranker 80 to promoter 82, and from metadata quality improver 78 to promoter 82). Figure 11 is a flow diagram of an exemplary full-text relevancy ranking process in accordance with the present invention. Media metadata describing the semantics of the content are sorted and grouped into broad categories (e.g., who, what, where, when) in step 156. For example, artist of a streaming media file, type of streaming media, date the streaming media was created, and creation location of the streaming media. These broad categories are individually weighted along with technical parameters such as bit rate, duration, fidelity (audio sampling rate), etc., in step 158. A relevance score is calculated for each URL in accordance with associated weights in step 160. The relevancy score is based upon several weighting criteria, such as the number of times a query term occurs in the metadata (term frequency), the number of links to the referenced Web site, number of terms between query

terms in the text for the metadata, and the file type selected for a search (e.g., wav, MP3, ram, wma).

**[0063]** For example, suppose a user enters a search query comprising the terms “Mozart”, “Magic Flute”, and “Red”. The full-text relevancy ranker 80, knowing that Mozart is a name of a composer (encoded rule), semantically associates Mozart with the who category and looks for “Mozart” in a field designated as WhoCreation. Magic Flute is recognized as a music composition and is semantically associated with the What category and looked for in the Title field. Weights of greater value are assigned to terms that are associated with semantic categories than to terms that are not associated with semantic categories. Thus, matches to “Mozart” and “Magic Flute” are assigned a greater weight, and accordingly a higher relevancy score, than the unrelated term “Red”. The full-text relevancy ranker 80 also considers technical parameters in the calculation of the relevancy score. In the current example, if the term news were added to the search query, the full-text relevancy ranker 80 looks for news pieces about Mozart and the Magic Flute, rather than for a piece of music. In an exemplary embodiment of the invention, full-text relevancy ranker 80 searches for news articles by considering the duration of the indexed files. For example, the full-text relevancy ranker 80 knows that news pieces typically have a shorter duration than music files (an encoded rule). Accordingly, the full-text relevancy ranker 80 assigns a higher score to files with shorter lengths. If full-text ranker 80 is incorporated as part of the workflow architecture, the database is updated with the full-text relevancy ranked data and the associated ids are queued and available to all agents. If full-text relevancy ranker 80 is not incorporated as part of the workflow architecture, no associated ids are queued and made available to all agents. Rather, the results are made directly available to search systems and/or users.

[0064] Referring again to Figure 3, the Promoter 82, formats and prioritizes data for a target search system (e.g., search engine). Promoter 82 adds, deletes, and/or updates the data (including metadata) associated with a media file in accordance with the requirements of the target search system.

5 Promoter 82 also provides an indication to the search system of the trustworthiness of the media data. In an exemplary embodiment of the system, trustworthiness is determined in accordance with predetermined encoded rules. For example, promoter 82 may determine that metadata associated with the title fields is the most trustworthy, and that metadata associated with the genre  
10 fields is less trustworthy. This hierarchy of trustworthiness is provided to the target search system in a compatible format. The target search system may then use this hierarchy of trustworthiness to conduct its search or pass the URLs on to the user.

[0065] The present invention may be embodied in the form of computer-  
15 implemented processes and apparatus for practicing those processes. The present invention may also be embodied in the form of computer program code embodied in tangible media, such as floppy diskettes, read only memories (ROMs), CD-ROMs, hard drives, high density disk, or any other computer-readable storage medium, wherein, when the computer program code  
20 is loaded into and executed by a computer, the computer becomes an apparatus for practicing the invention. The present invention may also be embodied in the form of computer program code, for example, whether stored in a storage medium, loaded into and/or executed by a computer, or transmitted over some transmission medium, such as over electrical wiring or cabling, through fiber  
25 optics, or via electromagnetic radiation, wherein, when the computer program code is loaded into and executed by a computer, the computer becomes an apparatus for practicing the invention. When implemented on a

